The Project and Data Management Plan

Project Title: Intent Recognition Chatbot with Sentiment Analysis

Research Question: In order to improve user experience, how can we create and deploy an intelligent chatbot that correctly understands user intent and reacts accordingly, while also utilizing sentiment an alysis?

Aims and Objectives:
The objective of this research is to improve natural language comprehension in airline travel queries b y creating a machine learning model for intent recognition using the ATIS dataset. Objectives include preprocessing the data, exploring various classification algorithms, evaluating model performance through defined metrics, and creating a prototype application for real-time user interaction. Documentation of methods and findings will ensure clarity and provide insights into user behaviour patterns while maintaining privacy standards. The models we will be comparing for intent classification will be:

- **Support Vector Machines (SVM)**: Good for high-dimensional spaces and effective when there is clear margin of separation.
- **Random Forests**: An ensemble method that can handle overfitting well due to its use of multiple decision trees.
- **Neural Network Model (LSTM):** Better in capturing context, this model is especially well-suited for sequential data, such as text..

Background: The emergence of conversational agents, or chatbots, has revolutionized how people utilize technology in a number of fields, including travel by air. Through the interpretation of natural language input and the production of contextually relevant responses, these AI-driven solutions are intended to enable smooth communication between users and service providers. Intent detection is a crucial element in improving user experience because of the complexity and variety of customer queries in the travel industry, which range from questions regarding luggage restrictions to booking and canceling flights (Joulin et al., 2017). With advancements in machine learning techniques, particularly in natural language processing (NLP), researchers have increasingly explored methods to classify user intents more effectively. One significant dataset for this purpose is the Airline Travel Information System (ATIS) dataset, which comprises thousands of annotated queries related to airline travel (Hemphill et al., 1990). Studies have shown that leveraging machine learning models can improve intent classification accuracy. For example, because they can identify sequential dependencies in text data, recurrent neural networks (RNNs) and long short-term memory networks (LSTMs) have shown encouraging results (Hochreiter & Schmidhuber, 1997). Furthermore, ensemble techniques like Random Forests have been popular because to their capacity to withstand overfitting and offer interpretability through the use of many decision trees (Breiman, 2001). The significance of creating chatbots that can both comprehend user intents and dynamically produce pertinent responses depending on those intents has been emphasized by recent research. This feature is necessary to keep the discussion flowing in an interesting way and guarantee user satisfaction during the chatbot system's engagement (Zhang et al., 2020). As we continue developing our airline-focused chatbot

model for intent recognition, we draw upon these foundational concepts from previous research while seeking innovative ways to enhance both functionality and usability.

Data: The project will utilize the Airline Travel Information System (ATIS) dataset, a benchmark for intent classification and slot filling in natural language processing. Comprising around 5,000 user queries related to airline travel—such as flight information, reservations, and cancellations—this dataset is annotated with intent labels and slot tags. Accessible through platforms like the Natural Language Toolkit (NLTK), it aids in training models to understand user intents and provide insights into various airline-related requests. Overall, the ATIS dataset is crucial for developing effective conversational agents in the aviation sector.

In terms of data distribution, the ATIS dataset features a variety of intents such as booking flights (`book_flight`), checking flight status (`flight_status`), changing itineraries (`change_flight`), and more specific actions like getting information about fare rules (`fare_info`). Each query has been labeled according to these intents along with relevant slots such as departure cities, arrival cities, dates, times, etc., enabling comprehensive training on both aspects of dialogue interaction. With its size and diverse range of queries representative of real-world scenarios faced by airline customers, the ATIS dataset stands out as an essential tool in developing robust conversational agents capable of efficiently managing customer interactions within the aviation sector.

Data Ethics:
For the project utilizing the Airline Travel Information System (ATIS) dataset, it is important to address several ethical considerations:

1. Data Source and Anonymization: No personally identifiable information (PII) is present in the ATIS dataset, which is made up of user queries on airline travel. Since the information has been gathered and disseminated for research purposes, there are no ethical issues with the use of personal data or the requirement for anonymization.

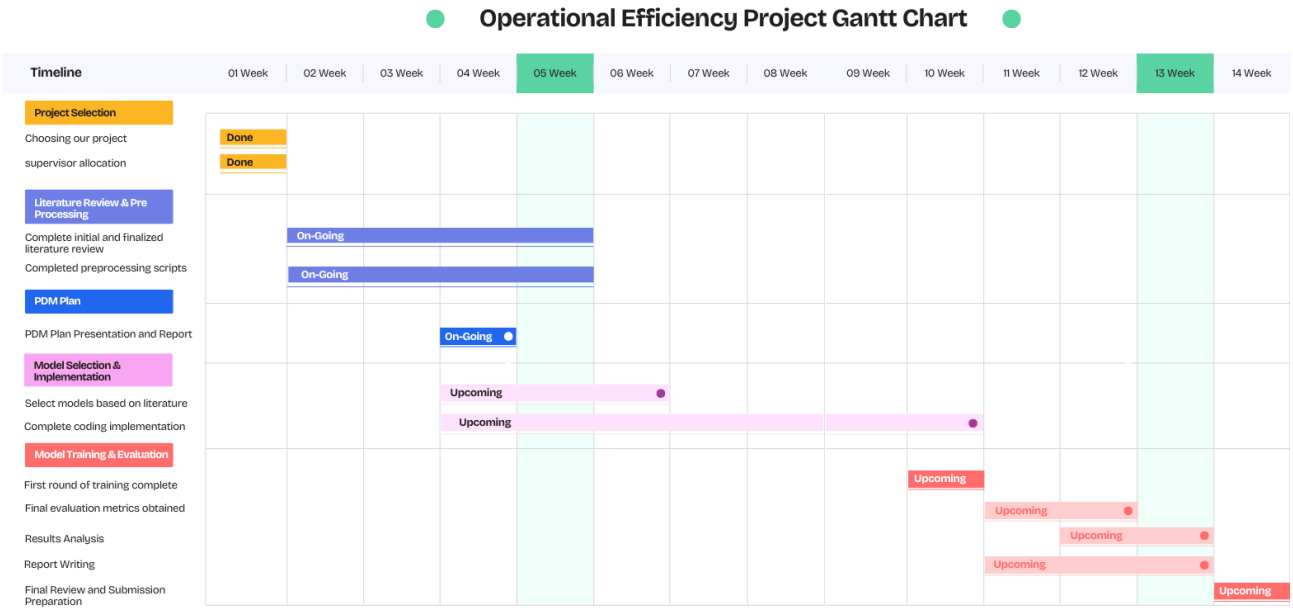
2. Data Usage: Our project will focus on training machine learning models for intent classification and slot filling using this publicly available dataset. We will ensure that our use of this data aligns with its intended research applications, adhering to academic integrity guidelines.

3. Involvement of Others: Since this project primarily involves working with an existing dataset rather than collecting new data from individuals, there are no additional participants involved who would require ethical approval from the University of Hertfordshire Ethics Committee.

4. Testing Code: Code testing will be conducted using standard practices such as unit tests and cross-validation techniques on the ATIS dataset itself, ensuring all evaluations are performed in a controlled manner without involving external human subjects.

Project Plan: The project plan outlines important tasks and a timeline for completion. First, I will conduct a literature review on machine learning in intent classification and slot filling. Next, I will

clean and prepare the ATIS dataset for analysis. Then, I will evaluate various machine learning models and choose several to implement using Python libraries. After that, I'll train these models, assess their performance, and analyse the results with accuracy metrics and visual aids. Finally, I'll document everything I've done and prepare for a final review before submitting the project according to university standards.



Data Management Plan: The project's data collecting, storage, backup, and version control methods are described in this data management plan. Data will be collected from public datasets in CSV format and stored locally and on cloud services like Google Drive. Regular backups will occur weekly locally and bi-weekly in the cloud. GitHub will be used for version control, with a minimum of weekly commits to track code development effectively. The plan ensures organized management of data and code throughout the project's lifecycle.

Reference List:

- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32.
- Hemphill, C., Godfrey Jr., J.J., & Doddington G.R. (1990). The ATIS Spoken Language Systems Pilot Corpus. *Proceedings of the workshop on Speech and Natural Language*.
- Hochreiter S., Schmidhuber J.(1997). Long short-term memory. *Neural Computation*, 9(8):1735–80.
- Joulin A., Mikolov T., Grave E., Bojanowski P ., Mikolov P.(2017). Bag of Tricks for Efficient Text Classification. *arXiv preprint arXiv:1607.01759*.
- Zhang Y., Zhao J., *Saleh M.*, & Liu Y.*(2020). Dialogpt: Large-scale generative pre-training for conversational response generation.* arXiv preprint arXiv:1911.00536